Statistiques avec R

SCI 1018

Tests d'hypothèse sur deux groupes

Marc J. Mazerolle

Département des sciences du bois et de la forêt, Université Laval

Avec révisions mineures de

Élise Filotas, *Département science et technologie, Université TÉLUQ*, et

Marc-Olivier Martin-Guay, *Département des sciences biologiques, Université du Québec à Montréal*

Table des matières

1 Introduction

Dans la leçon précédente, nous avons discuté de la démarche scientifique et de l'élaboration d'hypothèses scientifiques et statistiques. Deux types d'erreurs associées à la réalisation d'un test d'hypothèse ont été présentées : l'erreur de type I et l'erreur de type II. Le seuil de signification (α) correspond à la probabilité de rejeter une hypothèse nulle qui est vraie (erreur de type I), alors que la probabilité de ne pas rejeter une hypothèse nulle qui aurait dû être rejetée est donnée par β (erreur de type II). Nous avons vu le concept de la puissance statistique. Nous avons réalisé un exemple de test d'hypothèse statistique sur un échantillon (groupe) concernant la moyenne à l'aide du test t. Ce test peut être unilatéral ou bilatéral. Le test unilatéral cible une seule des deux queues de la distribution alors que le test bilatérale cible les deux queues de la distribution.

2 Tests d'hypothèses sur la moyenne de deux groupes indépendants

Comme nous l'avons vu précédemment, le but du test d'hypothèse sur un groupe est de déterminer si la moyenne d'une population diffère d'une valeur d'intérêt (test bilatéral) ou si la moyenne est inférieure ou supérieure à une valeur d'intérêt (test unilatéral). Ce même concept s'étend naturellement en présence de deux échantillons ou deux groupes, comme l'illustre l'exemple qui suit.

Exemple 4.1 On veut déterminer si les notes des élèves d'une école primaire sont différentes lorsqu'ils ont eu de l'aide aux devoirs pendant l'année comparativement à d'autres élèves n'ayant pas eu cette aide. Pour ce faire, on sélectionne aléatoirement 30 élèves qui recevront l'aide et 30 autres qui ne recevront pas l'aide. Les élèves auront le même enseignant et leurs capacités seront mesurées

dans un examen ministériel uniforme valant 40 % (tableau 1) (tableau 1).

$H_0 : \mu_{\text{aide}} = \mu_{\text{témoins}}$ (test bilatéral)

$H_a : \mu_{\text{aide}} \neq \mu_{\text{témoins}}$

$\alpha = 0.05$

Tableau 1 – Notes sur 40 des élèves dans les deux groupes de l'expérience.

Aide aux devoirs			Témoins		
36.36	36.10	34.97	19.20	19.31	21.06
37.70	37.22	39.05	21.21	19.41	18.89
35.75	36.32	35.42	18.78	17.66	19.24
35.92	37.55	35.08	20.00	18.21	19.92
36.24	36.90	36.92	20.73	19.88	18.97
36.04	37.86	37.28	19.50	20.88	20.96
37.17	36.52	35.60	20.18	20.15	18.97
35.65	36.03	36.00	20.06	21.88	21.26
37.57	34.66	36.26	20.12	19.21	19.99
36.49	36.22	37.67	19.98	20.76	21.23

On obtient les statistiques descriptives suivantes :

$$\bar{x}_{\text{aide}} = 36.48 \qquad \bar{x}_{\text{témoins}} = 19.92$$

$$s^2_{\text{aide}} = 0.96 \qquad s^2_{\text{témoins}} = 0.98$$

$$n_{\text{aide}} = 30 \qquad n_{\text{témoins}} = 30$$

On peut utiliser un test t pour déterminer s'il y a une différence entre les moyennes des deux groupes.

Dans le cas du test t sur un seul groupe, nous avons l'équation

$$t = \frac{\bar{x} - \mu}{SE}$$

où $\bar{x}$ correspond à la moyenne du groupe et μ représente la valeur à laquelle on désire comparer la moyenne. Comme nous l'avons vu à la leçon précédente, le test t sur un groupe suppose la normalité des résidus et l'indépendance des observations. En présence de deux groupes, le test t se calcule comme suit :

$$t = \frac{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}}{SE_{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}}}$$

Ici, les moyennes des deux groupes ($\bar{x}_{\text{groupe 1}}$, $\bar{x}_{\text{groupe 2}}$) sont comparées entre elles. Le dénominateur, $SE_{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}}$ correspond à l'erreur-type de la différence des moyennes. Voyons maintenant comment calculer cette valeur.

La variance de la différence des moyennes équivaut à la somme des variances séparées :

$$\sigma^2_{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}} = \sigma^2_{\text{groupe 1}} + \sigma^2_{\text{groupe 2}}$$

En plus des suppositions de normalité des résidus et de l'indépendance des observations, le test t sur deux groupes indépendants requiert que les variances des deux groupes soient égales ($\sigma^2_{\text{groupe 1}} = \sigma^2_{\text{groupe 2}}$). On parle alors d'homogénéité des variances. Il faut donc estimer la variance commune aux deux groupes. La meilleure estimation de cette variance commune consiste à calculer une variance combinée ou « poolée » (*pooled variance*) :

$$s^2_{\text{combinée}} = \frac{(n_{\text{groupe 1}} - 1)s^2_{\text{groupe 1}} + (n_{\text{groupe 2}} - 1)s^2_{\text{groupe 2}}}{n_{\text{groupe 1}} + n_{\text{groupe 2}} - 2} \,.$$

On utilise cette variance pour calculer l'erreur-type des différences des moyennes :

$$SE_{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}} = \sqrt{\frac{s^2_{\text{combinée}}}{n_{\text{groupe 1}}} + \frac{s^2_{\text{combinée}}}{n_{\text{groupe 2}}}} \,.$$

Maintenant, reprenons notre exemple sur le succès scolaire des élèves en considérant la variance combinée.

Exemple 4.2. On peut effectuer un test t pour comparer les deux groupes d'élèves de l'exemple 4.1. Commençons par calculer la variance combinée qui sera nécessaire pour obtenir le dénominateur du test t sur deux groupes :

$$\begin{aligned} s^2_{\text{combinée}} &= \frac{(n_{\text{aide}}-1)s^2_{\text{aide}} + (n_{\text{témoins}}-1)s^2_{\text{témoins}}}{n_{\text{aide}} + n_{\text{témoins}} - 2} \\ s^2_{\text{combinée}} &= \frac{(30-1)\cdot 0.96 + (30-1)\cdot 0.98}{30+30-2} \\ s^2_{\text{combinée}} &= 0.97 \end{aligned}$$

Ensuite calculons l'erreur-type sur la différence des moyennes :

$$\begin{aligned} SE_{\bar{x}_{\text{aide}}-\bar{x}_{\text{témoins}}} &= \sqrt{\frac{s^2_{\text{combinée}}}{n_{\text{aide}}} + \frac{s^2_{\text{combinée}}}{n_{\text{témoins}}}} \\ SE_{\bar{x}_{\text{aide}}-\bar{x}_{\text{témoins}}} &= \sqrt{\frac{0.97}{30} + \frac{0.97}{30}} \\ SE_{\bar{x}_{\text{aide}}-\bar{x}_{\text{témoins}}} &= 0.25 \end{aligned}$$

La statistique du test s'obtient ainsi :

$$\begin{aligned} t &= \frac{\bar{x}_{\text{aide}} - \bar{x}_{\text{témoins}}}{SE_{\bar{x}_{\text{aide}}-\bar{x}_{\text{témoins}}}} \\ t &= \frac{36.48 - 19.92}{0.25} \\ t &= 65.217 \end{aligned}$$

Nous déterminons la probabilité d'observer une valeur $t = 65.217$ dans des populations où H_0 est vraie, soit l'énoncé $P(|t_{df=58}| \geq 65.217)$. On peut consulter la distribution du t de Student avec 58 degrés de liberté ($df = n_{\text{aide}} + n_{\text{témoins}} - 2$) dans un livre de statistique ou encore avec `R` à l'aide de `pt( )`. On constate que la valeur observée $t = 65.217$ est très peu probable lorsque H_0 est vraie, puisque $P(|t_{df=58}| \geq 65.217) < 0.0001$. On rejette donc H_0 et on conclut que les deux moyennes diffèrent, c'est-à-dire que la moyenne des notes des élèves ayant eu

l'aide aux devoirs diffère de celle des élèves n'ayant pas eu cette aide.

Dans R, il est bien sûr possible de réaliser le tout plus succinctement. Le jeu de données est stocké dans le fichier `eleves.txt`. Il faudra importer ce fichier avant de pouvoir accéder aux données :

```
> ##importer le jeu de données
> eleves <- read.table("eleves.txt", header = TRUE)
> ##premières observations
> head(eleves)
   Note Type
1 36.36 aide
2 37.70 aide
3 35.75 aide
4 35.92 aide
5 36.24 aide
6 36.04 aide
```

Tout comme pour le test t sur un groupe, la fonction `t.test( )` permet de réaliser le test t sur deux groupes indépendants. Dans notre cas, nous utiliserons des arguments additionnels, notamment `var.equal = TRUE`, qui spécifie la supposition de variances homogènes entre les deux groupes. De plus, nous pourrons aussi utiliser une syntaxe sous forme de formule (c.-à-d., `Note` $\sim$ `Type`) pour indiquer la variable réponse (`Note`), la variable qui distingue les élèves avec aide aux devoirs et les témoins (`Type`) et l'argument `data`. L'argument `data` permet d'indiquer à R dans quel objet se trouvent les variables employées dans l'analyse. En présence de cet argument, il n'est pas nécessaire d'utiliser le symbole `$` à l'intérieur de la fonction `t.test( )` – `eleves$Note` $\sim$ `eleves$Type` donnera la même chose

que `Note ~ Type, data = eleves`. Cette syntaxe est utilisée dans la plupart des analyses statistiques de R.

```
> ##on exécute le test t pour deux groupes
> t.test(Note ~ Type, data = eleves, var.equal = TRUE)

        Two Sample t-test


data:  Note by Type
t = 65.217, df = 58, p-value < 2.2e-16
alternative hypothesis: true difference in means between group aide and group temoins
95 percent confidence interval:
 16.0556 17.0724
sample estimates:
   mean in group aide mean in group temoins
                36.484                19.920
```

On remarque que `test.t( )` donne le test t bilatéral par défaut, comme c'est le cas pour le test t sur un groupe. Nous pouvons facilement choisir un test unilatéral en modifiant l'argument `alternative`. Nous verrons un exemple de test unilatéral plus loin dans ce document (exemple 4.4).

À noter que la probabilité retournée par `t.test( )` est très faible (`p.value < 2.2e-16`). Par convention, on indique dans un rapport présentant des résultats statistiques que la probabilité est inférieure à une petite valeur, telle que $P < 0.0001$ ou $P < 0.001$. Même si la probabilité est très faible, on n'indique jamais que $P = 0$. Il est erroné d'écrire $P = 0$ puisque cela équivaut à dire que l'événement est impossible. Il faut distinguer entre **impossible** (ne se produira jamais) et **peu probable** (peut se produire avec une très faible probabilité). Par analogie, la probabilité de gagner à une loterie est peut-être faible ($P < 0.0001$) mais non nulle si on a acheté un billet. Par contre, si on n'achète aucun billet, on n'a aucune chance de gagner et, dans ce cas, on peut écrire $P = 0$.

2.1 Suppositions

Les suppositions ou conditions à respecter pour le test t à deux groupes sont les suivantes :

- l'indépendance des observations : les deux échantillons sont indépendants et les observations proviennent d'un échantillonnage aléatoire ;
- la normalité des résidus : les deux groupes proviennent de populations normales et leurs erreurs suivent une distribution normale ;
- l'homogénéité des variances : les deux groupes ont la même variance.

2.1.1 Indépendance des observations

Le meilleur moyen de respecter cette condition est d'utiliser un design complètement aléatoire afin de récolter les données. Comme nous l'avons vu dans la leçon précédente, des mesures répétées sur les mêmes unités avant et après un traitement ne constituent pas des observations indépendantes.

2.1.2 Normalité des résidus

Afin de vérifier la normalité des résidus, nous pouvons utiliser les mêmes diagnostics que ceux présentés pour le test t à un groupe. Les tests formels ou les méthodes graphiques peuvent nous aider à détecter des déviations de cette condition. Dans notre exemple sur le succès scolaire des élèves, nous pourrions procéder comme suit :

```
> ##on crée un sous-jeu de données pour les élèves ayant eu l'aide aux devoirs
> eleves.aide <- eleves[eleves$Type == "aide", "Note"]
> eleves.aide
 [1] 36.36 37.70 35.75 35.92 36.24 36.04 37.17 35.65 37.57
[10] 36.49 36.10 37.22 36.32 37.55 36.90 37.86 36.52 36.03
[19] 34.66 36.22 34.97 39.05 35.42 35.08 36.92 37.28 35.60
[28] 36.00 36.26 37.67
```

```
> ##on crée un sous-jeu de données les élèves témoins
> eleves.temoins <- eleves[eleves$Type == "temoins", "Note"]
> eleves.temoins
 [1] 19.20 21.21 18.78 20.00 20.73 19.50 20.18 20.06 20.12
[10] 19.98 19.31 19.41 17.66 18.21 19.88 20.88 20.15 21.88
[19] 19.21 20.76 21.06 18.89 19.24 19.92 18.97 20.96 18.97
[28] 21.26 19.99 21.23
> ##résidus du groupe avec aide aux devoir
> res.aide <- eleves.aide - mean(eleves.aide)
> ##résidus du groupe témoins
> res.temoins <- eleves.temoins - mean(eleves.temoins)
> ##on combine les résidus des deux groupes
> res.combo <- c(res.aide, res.temoins)
> ##test formel Anderson-Darling
> library(nortest)
> ad.test(res.combo)

        Anderson-Darling normality test

data:  res.combo
A = 0.3122, p-value = 0.5408
> ##graphique quantile-quantile
> qqnorm(res.combo, ylab = "Quantiles observées",
         xlab = "Quantiles théoriques",
         main = "Graphique quantile-quantile")
> qqline(res.combo)
```

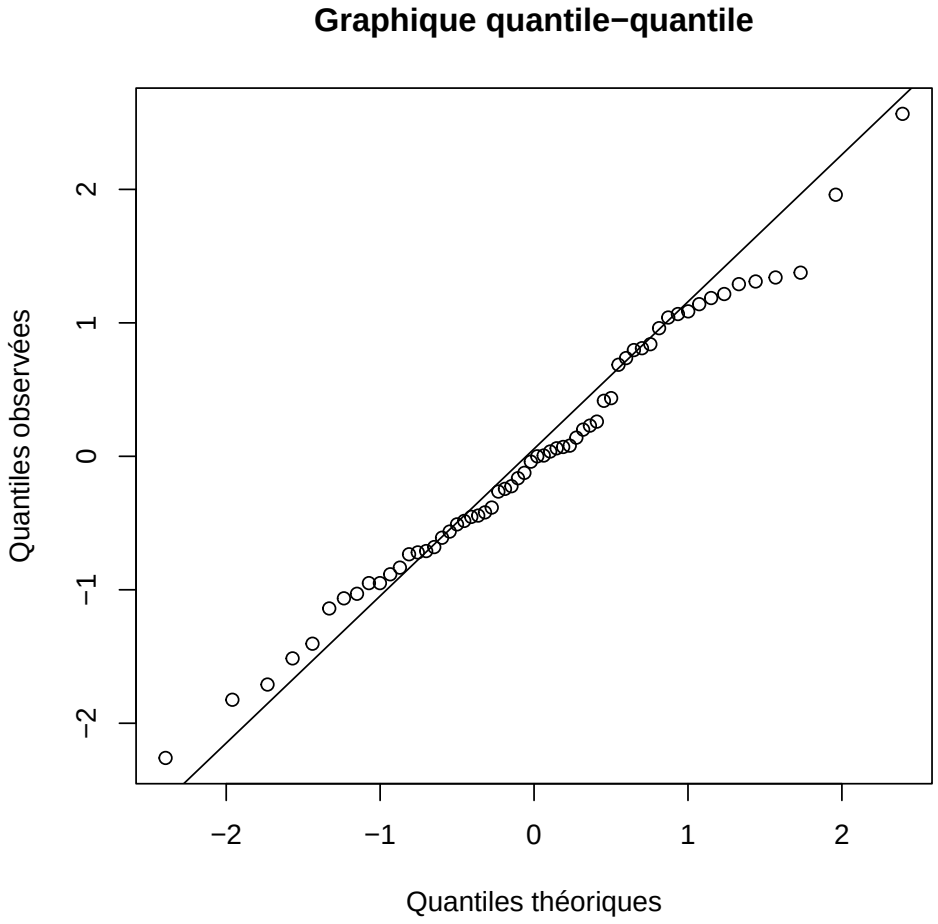

FIGURE 1 – Graphique quantile-quantile à partir des résidus du test t sur deux groupes indépendants effectué sur les données du succès scolaire des élèves.

Les diagnostics suggèrent que la supposition de normalité des résidus est bien respectée (fig. 1). En effet, le résultat du test de Anderson-Darling (`p.value = 0.5408`) ne permet pas de rejeter l'hypothèse nulle selon laquelle les résidus suivent une distribution normale avec un seuil $\alpha = 0.05$. De plus, les quantiles des observations s'alignent bien avec les quantiles théoriques d'une distribution normale.

2.1.3 Homogénéité de la variance

Le test t sur deux groupes indépendants suppose que les deux groupes comparés ont des variances égales. C'est ce qu'on appelle la supposition d'**homogénéité de la variance**,

aussi appelée **homoscédasticité** (*homogeneity of variance*, *homoscedasticity*, *homoskedasticity*). Lorsque deux groupes ont des variances inégales, nous dirons que l'homogénéité de la variance n'est pas respectée, ou de façon équivalente, que les variances sont hétérogènes (*heterogeneous variances*, *heteroscedasticity*, *heteroskedasticity*). Cette supposition est très importante et il est primordial de la vérifier. On peut évaluer si cette supposition est valide à l'aide de méthodes formelles telles que le test de Fligner-Killeen (`fligner.test( )`), de Bartlett (`bartlett.test( )`) ou celui de Levene (`leveneTest( )`) du package `car`. Toutefois, ces tests formels peuvent être sensibles à la non-normalité, c'est-à-dire que le rejet de l'hypothèse nulle (H_0 : homogénéité de la variance) peut être dû à des déviations de la normalité plutôt qu'à l'hétérogénéité de la variance. Ainsi, nous utiliserons plutôt des méthodes graphiques pour diagnostiquer les problèmes d'hétérogénéité de la variance.

La méthode graphique la plus simple consiste à observer le patron de la variance à l'aide des résidus. Puisque nous n'avons que deux groupes, nous pouvons utiliser un diagramme de boîtes et moustaches (*boxplot*). Si la condition d'homogénéité de la variance est respectée, nous observerons deux boîtes de même hauteur sur le graphique.

```
> ##on met les résidus dans le jeu de données
> eleves$res.combo <- res.combo
> ##on crée le graphique
> boxplot(res.combo ~ Type, data = eleves)
```

On voit que les boîtes des deux groupes sont très semblables (la distance interquartile est similaire). On peut donc procéder avec le test t (fig. 2). Dans d'autres cas, les variances des deux groupes sont différentes et la supposition d'homogénéité de la variance n'est pas respectée (fig. 3). Lorsqu'il y a un doute sur l'interprétation du diagramme, il vaut mieux utiliser des tests qui ne nécessitent pas l'homogénéité de la variance (voir la prochaine section).

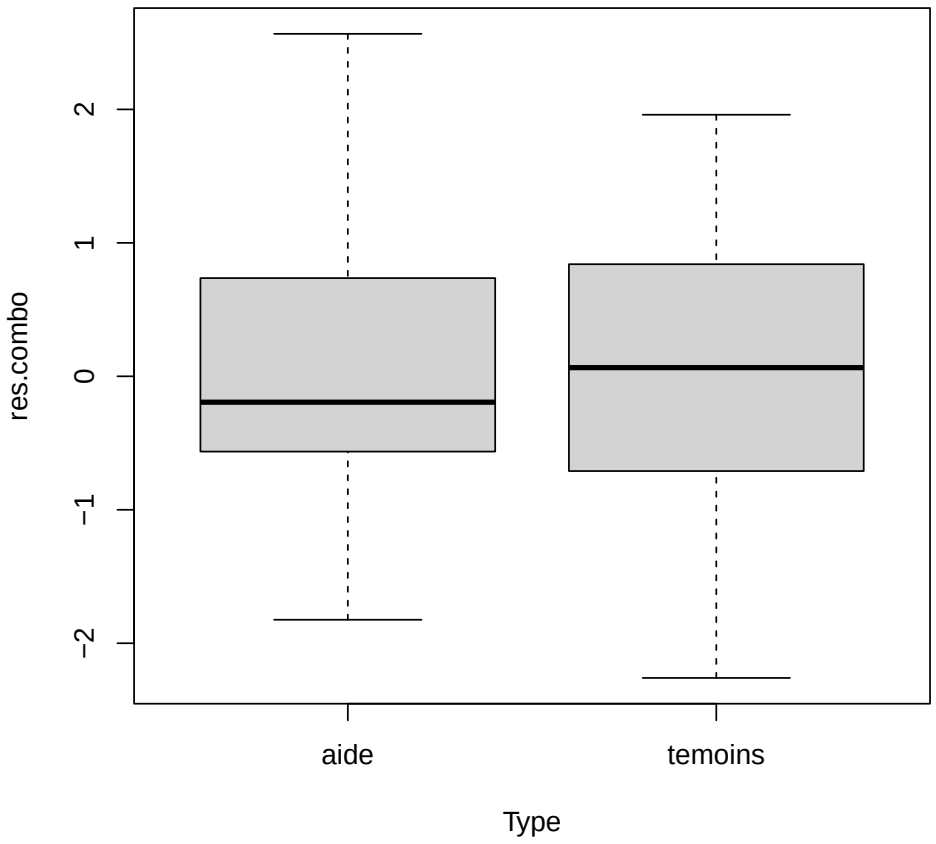

FIGURE 2 – Diagramme de boîtes et moustaches à partir des résidus du test t sur deux groupes indépendants effectué sur les données du succès scolaire des élèves.

2.1.4 Robustesse du test t

Le test t est robuste aux déviations, particulièrement si les deux groupes sont de tailles identiques et si on teste des hypothèses bilatérales. La robustesse du test augmente avec la taille de l'échantillon. Lorsque la supposition d'homogénéité de la variance ne peut être respectée, une version modifiée du test t peut être utilisée, le test t de Welch (*Welch's t-test*). C'est d'ailleurs ce test t qui est exécuté par défaut, avec `t.test( )` (c.-à-d., l'argument `var.equal` prend la valeur `FALSE` par défaut). L'exemple suivant illustre ce qu'on obtient avec cette modification du test t.

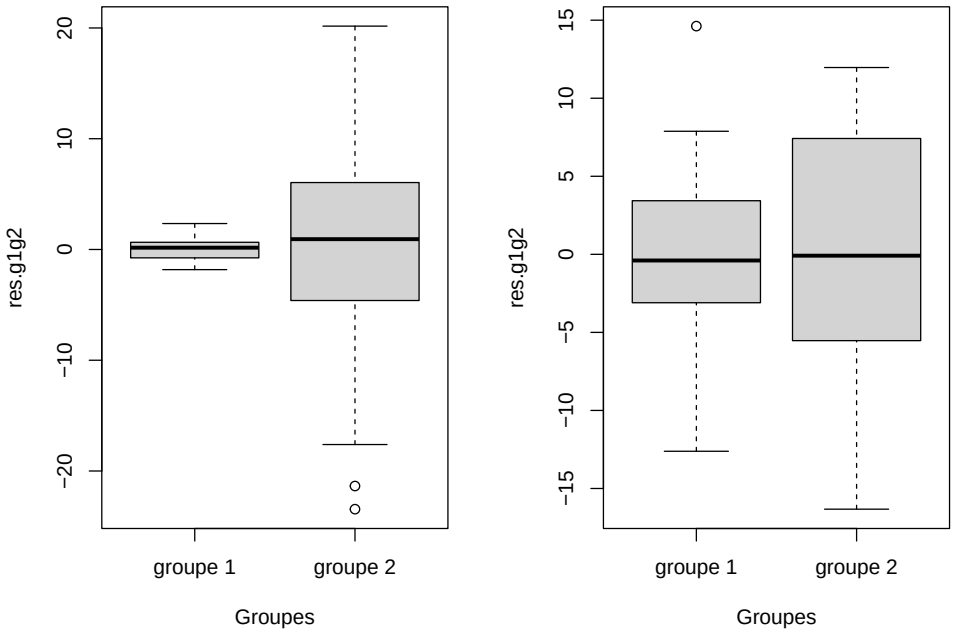

FIGURE 3 – Diagrammes de boîtes et moustaches sur les résidus de groupes avec variances inégales. On remarque que les boîtes ont des hauteurs différentes (distance interquartile différente). Dans chaque diagramme, le groupe 1 varie moins que le groupe 2.

Exemple 4.3 Le test t de Welch est une version modifiée du test t classique et est adapté aux situations de variances hétérogènes. Alors que le test t utilise un terme de variance combinée, le test t de Welch inclut les variances des deux groupes dans le calcul de l'erreur-type sur la différence des moyennes :

$$SE_{\bar{x}_{\text{groupe 1}} - \bar{x}_{\text{groupe 2}}} = \sqrt{\frac{s^2_{\text{groupe 1}}}{n_{\text{groupe 1}}} + \frac{s^2_{\text{groupe 2}}}{n_{\text{groupe 2}}}}$$

Le calcul des degrés de liberté est plus complexe avec le test t de Welch :

$$df = \frac{\left(\frac{s^2_{\text{groupe 1}}}{n_{\text{groupe 1}}} + \frac{s^2_{\text{groupe 2}}}{n_{\text{groupe 2}}}\right)}{\frac{\left(\frac{s^2_{\text{groupe 1}}}{n_{\text{groupe 1}}}\right)^2}{n_{\text{groupe 1}}-1} + \frac{\left(\frac{s^2_{\text{groupe 2}}}{n_{\text{groupe 2}}}\right)^2}{n_{\text{groupe 2}}-1}}.$$

On peut comparer la différence entre les résultats du test t de Student et ceux du test t de Welch concernant les données relatives au succès scolaire des élèves :

```
> ##test t de Student
> t.test(Note ~ Type, data = eleves, var.equal = TRUE)

        Two Sample t-test


data:  Note by Type
t = 65.217, df = 58, p-value < 2.2e-16
alternative hypothesis: true difference in means between group aide and group te
95 percent confidence interval:
 16.0556 17.0724
sample estimates:
   mean in group aide mean in group temoins
               36.484                19.920
> ##test t de Welch
> t.test(Note ~ Type, data = eleves, var.equal = FALSE)

        Welch Two Sample t-test


data:  Note by Type
t = 65.217, df = 57.993, p-value < 2.2e-16
alternative hypothesis: true difference in means between group aide and group te
95 percent confidence interval:
 16.0556 17.0724
sample estimates:
   mean in group aide mean in group temoins
               36.484                19.920
```

On constate que les résultats des tests sont pratiquement les mêmes puisque la

supposition d'homogénéité de la variance est respectée. Seuls les degrés de libertés diffèrent légèrement.

Exemple 4.4 On effectue une expérience sur le temps passé à regarder son téléphone portable pour des étudiants de 19 à 23 ans dans deux situations distinctes : dîner avec des étudiants du même département et dîner avec des étudiants de départements différents. On recrute 30 volontaires que l'on assigne aléatoirement aux deux situations, à raison de 15 individus dînant avec des compères et 15 individus dînant avec des étudiants inconnus d'une autre discipline. Le temps que chaque volontaire passe à regarder son téléphone pendant le dîner de 1h30 est noté. À la fin de l'expérience, on obtient les résultats du tableau 2.

Tableau 2 – Temps en minutes passé sur son téléphone portable de différents individus dînant avec des compères ou des étudiants inconnus de d'autres disciplines.

Inconnus			Compères		
27.05	13.00	23.52	20.39	35.88	12.49
14.98	5.15	4.43	13.76	35.78	33.20
26.19	27.65	9.39	30.06	24.73	26.80
11.92	18.46	19.88	34.21	42.14	27.49
16.49	5.27	13.40	23.96	18.37	40.78

Les données se trouvent dans le fichier `temps.txt`. Puisqu'on s'attend à ce que les individus dînant avec des compères passent plus de temps sur leurs téléphones portables que ceux dînant avec des inconnus (car ils sont désinhibés en présence de leurs compères et moins polis), nous effectuerons un test unilatéral.

$H_0 : \mu_{\text{inconnus}} \geq \mu_{\text{compères}}$ (test unilatéral)

$H_a : \mu_{\text{inconnus}} < \mu_{\text{compères}}$

$\alpha = 0.05$

On importe le jeu de données :

```
> temps <- read.table("temps.txt", header = TRUE)
> ##premières observations
> head(temps)
     Temps Accompagnateurs
1 27.05083        inconnus
2 14.98468        inconnus
3 26.18627        inconnus
4 11.92027        inconnus
5 16.49150        inconnus
6 13.00014        inconnus
> ##dernières observations
> tail(temps)
      Temps Accompagnateurs
25 18.36979         comperes
26 12.49332         comperes
27 33.20128         comperes
28 26.80184         comperes
29 27.49096         comperes
30 40.77970         comperes
```

On peut effectuer le test t de Student pour tester notre hypothèse :

```
> ##temps pour dîner avec inconnus
> inconnus <- temps[temps$Accompagnateurs == "inconnus",  "Temps"]
> ##temps pour dîner avec compères
> comperes <- temps[temps$Accompagnateurs == "comperes", "Temps"]
> ##test unilatéral avec x plus grand que y
```

```
> t.out <- t.test(x = comperes, y= inconnus, data = temps,
                  var.equal = TRUE, alternative = "greater")
> t.out

        Two Sample t-test


data:  comperes and inconnus
t = 3.8997, df = 28, p-value = 0.0002747
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 6.887848      Inf
sample estimates:
mean of x mean of y
 28.00328  15.78609
```

Toutefois, avant de se lancer dans l'interprétation de l'analyse, il faut vérifier les suppositions.

```
> ##calculer les résidus - groupe inconnus
> inconnus.res <- inconnus - mean(inconnus)
> ##calculer les résidus - groupe comperes
> comperes.res <- comperes - mean(comperes)
> ##combiner résidus dans jeu de données
> temps$Res <- c(inconnus.res, comperes.res)
> ##vérifier normalité avec test Anderson-Darling
> ad.test(temps$Res)

        Anderson-Darling normality test


data:  temps$Res
A = 0.24703, p-value = 0.732
```

```
> ##combiner graphique dans une fenêtre
> par(mfrow = c(1, 2))
> ##graphique quantile-quantile
> qqnorm(temps$Res, ylab = "Quantiles observés",
         xlab = "Quantiles théoriques",
         main = "Graphique quantile-quantile")
> qqline(temps$Res)
> ##ajouter lettre a
> text(x = -2, y = 14.5, labels = "a", cex = 1.2)
> ##diagramme de boîtes et moustaches - homoscédasticité
> boxplot(Res ~ Accompagnateurs, data = temps)
> ##ajouter lettre b
> text(x = 0.5, y = 14.5, labels = "b", cex = 1.2)
```

Le test de normalité d'Anderson-Darling suggère que la supposition de normalité est respectée et il en va de même avec le graphique quantile-quantile (fig. 4a). La figure 4b ne montre pas de signes d'hétérogénéité de la variance. On peut procéder à l'inteprétation des résultats. Puisque $P(t_{df=28}3.9) = 0.00027$, on rejette H_0 et on conclut que le temps passé à regarder son téléphone portable est plus grand lorsque les étudiants dînent avec leurs compères plutôt qu'avec des inconnus. Autrement dit, la valeur de t observée est beaucoup plus extrême que ce qui est attendu selon H_0. Finalement, on peut illustrer le résultat dans un graphique de la moyenne de chaque groupe accompagnée de son intervalle de confiance respectif (fig. 5).

```
> ##moyenne des groupes
> moy.inconnus <- mean(inconnus)
```

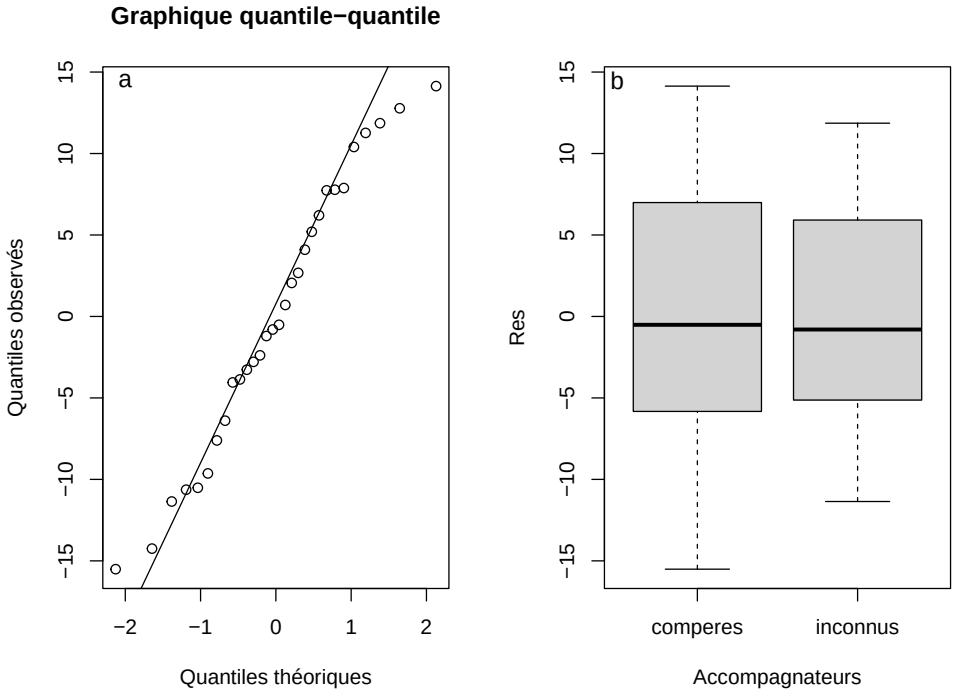

FIGURE 4 – Graphique quantile-quantile sur les résidus du test t sur deux groupes indépendants sur les données de temps passés par des étudiants sur leurs téléphones portables (a) et diagramme de boîtes et moustaches pour vérifier la supposition d'homogénéité de la variance à partir des résidus du test t sur deux groupes indépendants avec avec les mêmes données (b).

```
> moy.comperes <- mean(comperes)
> ##SD des groupes
> sd.inconnus <- sd(inconnus)
> sd.comperes <- sd(comperes)
> ##n des groupes
> n.inconnus <- length(inconnus)
> n.comperes <- length(comperes)
> ##SE des moyennes
> SE.inconnus <- sd.inconnus/sqrt(n.inconnus)
> SE.comperes <- sd.comperes/sqrt(n.comperes)
> ##IC's à 95%
```

```
> IC.inf95.inconnus <- moy.inconnus -
    qt(p = 0.025, df = n.inconnus - 1) * SE.inconnus
> IC.sup95.inconnus <- moy.inconnus +
    qt(p = 0.025, df = n.inconnus - 1) * SE.inconnus
> IC.inf95.comperes <- moy.comperes -
    qt(p = 0.025, df = n.comperes - 1) * SE.comperes
> IC.sup95.comperes <- moy.comperes +
    qt(p = 0.025, df = n.comperes - 1) * SE.comperes
> ##graphique
> plot(y = 0, x = 0, xlab = "Accompagnateurs",
       ylab = "Temps passé sur son téléphone (minutes)",
       main = "Moyennes ± IC à 95%",
       ylim = range(c(IC.inf95.inconnus, IC.inf95.comperes,
                      IC.sup95.inconnus, IC.sup95.comperes)),
       xlim = c(0, 1),
       xaxt = "n")
> ##ajout de l'axe des x
> axis(side = 1, at = c(0.3, 0.75), labels = c("Inconnus", "Compères"))
> ##ajout des points
> points(y = c(moy.inconnus, moy.comperes), x = c(0.3, 0.75))
> # ajout des barres d'erreur
> arrows(x0 = c(0.3, 0.75), x1 = c(0.3, 0.75),
         y0 = c(IC.inf95.inconnus, IC.inf95.comperes),
         y1 = c(IC.sup95.inconnus, IC.sup95.comperes),
         length = 0.05, code = 3, angle = 90)
```

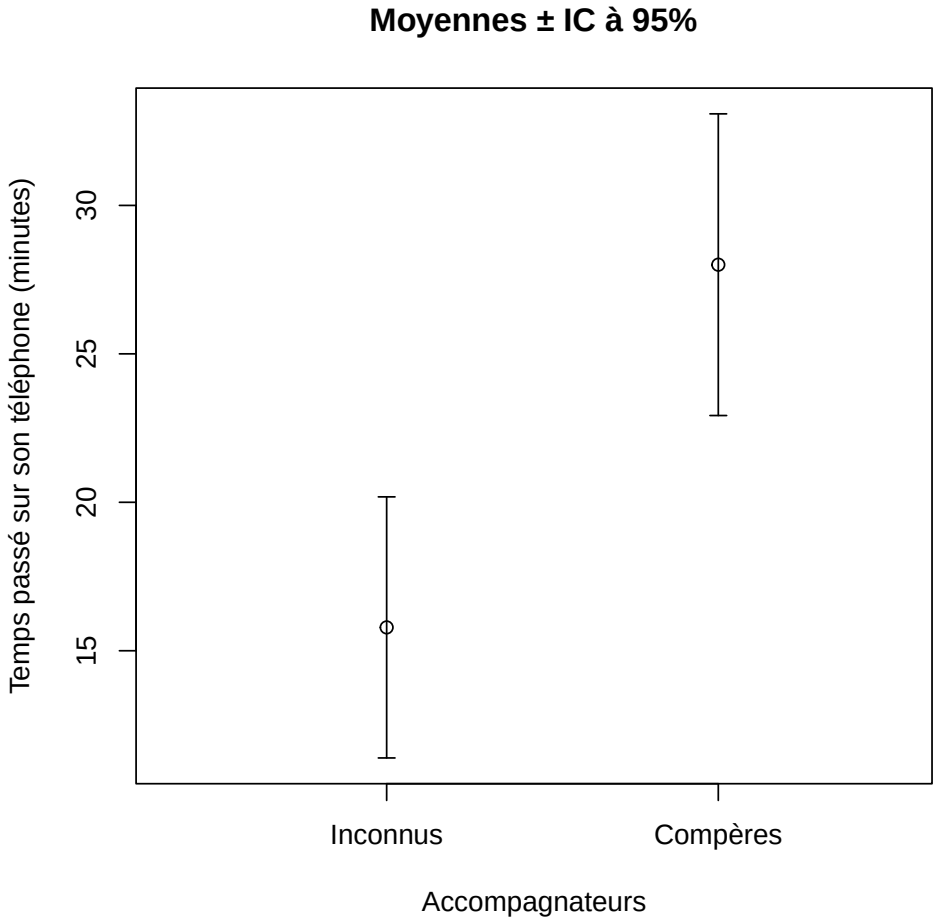

FIGURE 5 – Moyennes du temps passé sur leurs téléphones par des étudiants dînant accompagnés de compères et d'inconnus. Les barres d'erreurs correspondent à des intervalles de confiance à 95%.

3 Tests d'hypothèse pour deux groupes appariés

Dans certains dispositifs d'échantillonnage, on retrouve des données qui viennent en paires. On dira que les données sont **appariées**, c'est-à-dire que chaque observation d'un groupe est liée avec celle d'un autre groupe. Par exemple,

- on prend le pouls de jumeaux chez qui on a administré deux médicaments (un jumeau a le traitement 1, le deuxième a le traitement 2) ;
- on détermine l'asymétrie [1] chez une espèce de moustique et on mesure la longueur de

1. L'asymétrie (absence de symétrie) est une propriété de certains individus qui présentent des caractéris-

l'aile gauche et de l'aile droite. L'aile gauche d'un moustique constitue une observation du premier groupe alors que son aile droite est la donnée appariée qui fait partie du second groupe ;

- on mesure la croissance des mêmes arbres avant et après l'ajout d'un fertilisant. La première mesure de croissance d'un arbre constitue donc une donnée du premier groupe alors que la deuxième mesure du même arbre est la donnée appariée qui fait partie du second groupe.

Dans chaque exemple ci-dessus, on a des données sous forme de paires. Un dispositif avec des éléments appariés peut être une bonne stratégie pour réduire la variabilité de ce qu'on désire mesurer. Il réduit la variance entre deux observations d'une même paire. Ce dispositif est particulièrement utile lorsque la variabilité entre les individus (paires) est grande. On reconnaît que les paires diffèrent entre elles, mais la variabilité **intrapaire** (c.-à-d., à l'intérieur d'une même paire) est moindre.

On ne peut pas considérer des mesures appariées comme indépendantes les unes des autres et cette structure doit être prise en compte lors de l'analyse. En d'autres mots, il est inapprorié de considérer ces données comme provenant de deux groupes indépendants. Par conséquent, on ne peut pas faire abstraction de cette structure et utiliser le test t sur deux groupes qui sont indépendants. En présence de données appariées, nous utilisons plutôt le **test t pour données appariées** (*paired t-test*). Ce test est réalisé sur les différences entre les groupes et il correspond à un test t sur un groupe. Illustrons avec un exemple.

Exemple 4.5 Dans une étude d'impact environnementale, on vise à évaluer l'effet des eaux usées sur le nombre d'espèces de poissons dans différents cours d'eau. Pour ce faire, on sélectionne 16 cours d'eau dans lesquels se déversent des eaux usées. On échantillonne les poissons en amont (avant la décharge) et en aval (après

tiques différentes de part et d'autre d'un axe du corps. Si on mesure le bras gauche et le bras droit du même individu, on dira que la mesure est symétrique si elle est de la même valeur pour les deux membres.

la décharge) des sources d'eaux usées. Tous les cours d'eau sont visités à la même période de l'année. Les données se trouvent dans le fichier `poissons.txt`.

```
> poissons <- read.table("poissons.txt", header = TRUE)
> head(poissons)
  Aval Amont
1   20    23
2   15    16
3   10    10
4    5     4
5   20    22
6   15    15
> tail(poissons)
   Aval Amont
11   10    11
12    5     5
13   20    22
14   15    14
15   10    10
16    5     6
```

On peut utiliser un test t pour données appariées pour déterminer si le nombre d'espèces en aval est inférieure à celui en amont des décharges d'eaux usées. Dans cet exemple, nous définissons la différence du nombre d'espèces comme étant `Aval - Amont`.

$H_0 : \mu_{\text{différence}} \geq 0$ (test unilatéral)

$H_a : \mu_{\text{différence}} < 0$

$\alpha = 0.05$

À noter que le choix de calcul de la différence est totalement arbitraire et on aurait pu calculer plutôt `Amont - Aval`. Le cas échéant, il aurait fallu modifier H_0 à $\mu_{\text{différence}} \leq 0$ et H_a à $\mu_{\text{différence}} > 0$, puisqu'une valeur positive de la différence est conforme à notre prédiction que le nombre d'espèces est plus élevé en amont. Dans tous les cas, H_a doit être conforme à la prédiction associée à notre hypothèse scientifique. Le test t pour données appariées s'effectue sur la différence entre les paires d'observations de chaque groupe :

```
> poissons$Diff <- poissons$Aval - poissons$Amont
> head(poissons)
  Aval Amont Diff
1   20    23   -3
2   15    16   -1
3   10    10    0
4    5     4    1
5   20    22   -2
6   15    15    0

> mean.diff <- mean(poissons$Diff)
> mean.diff
[1] -0.875
> SD <- sd(poissons$Diff)
> SD
[1] 1.147461
> SE <- SD/sqrt(nrow(poissons))
> SE
[1] 0.2868652
```

On sait que :

$$\begin{aligned} \bar{x}_{\text{différence}} &= -0.9 \\ SD_{\text{différence}} &= 1.15 \\ SE_{\text{différence}} &= 0.29 \end{aligned}$$

On obtient ainsi :

$$\begin{aligned} t &= \frac{\bar{x}_{\text{différence}} - \mu_{H_0}}{SE_{\text{différence}}} \\ t &= \frac{-0.9 - 0}{0.29} \\ t &= -3.05 \end{aligned}$$

On peut effectuer rapidement le test dans `R` à l'aide de `t.test( )` :

```
> ##test t à un groupe sur la différence
> t.test(poissons$Diff, alternative = "less")

        One Sample t-test


data:  poissons$Diff
t = -3.0502, df = 15, p-value = 0.00405
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.3721108
sample estimates:
mean of x
   -0.875
```

Vous aurez remarqué que le test t pour données appariées n'est nul autre qu'un test t sur un seul groupe (voir Leçon 3). Ce groupe est constitué des différences

entre les paires d'observations. Le test t pour données appariées doit respecter les mêmes suppositions que celles du test t à un groupe, puisque le test t pour données appariées **est un test t effectué sur un seul groupe**. On peut aussi fournir les données brutes de chaque paire à la fonction `t.test( )`, mais il est très important d'inclure l'argument `paired = TRUE`.

```
> ##même test t, mais utilisant syntaxe de formule
> ##et l'argument paired = TRUE
> t.test(poissons$Aval, poissons$Amont, paired = TRUE,
         alternative = "less")

        Paired t-test


data:  poissons$Aval and poissons$Amont
t = -3.0502, df = 15, p-value = 0.00405
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.3721108
sample estimates:
mean of the differences
                 -0.875

> ##calcul des résidus
> res.diff <- poissons$Diff - mean(poissons$Diff)
> ##graphique quantile-quantile
> qqnorm(res.diff, ylab = "Quantiles observés",
         xlab = "Quantiles théoriques",
         main = "Graphique quantile-quantile")
> qqline(res.diff)
```

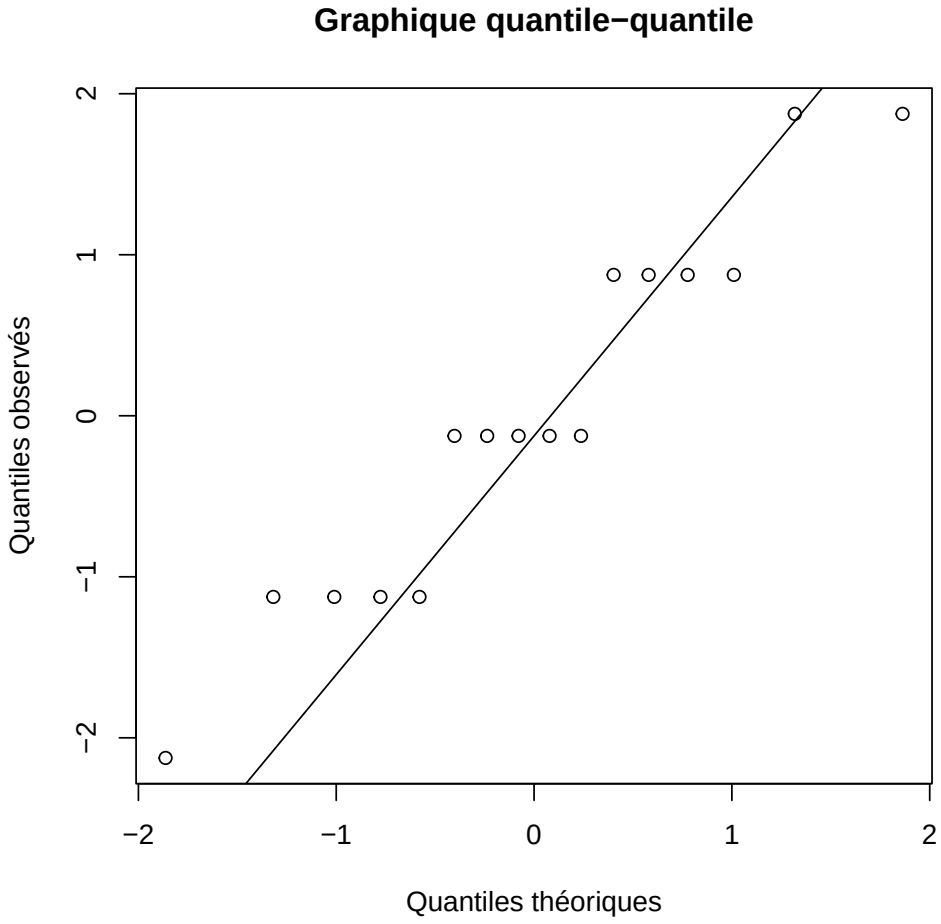

FIGURE 6 – Vérification de la normalité à partir des résidus des données appariées sur le nombre d'espèces de poissons.

La vérification de la normalité indique que les résidus ont une distribution normale (fig. 6). On conclut qu'il est peu probable d'observer une valeur t de -3.05 ou plus extrême (ici, plus faible étant donné notre hypothèse alternative) avec 15 degrés de liberté lorsque H_0 est vraie. Effectivement, $P(t_{15} \leq -3.05) = 0.004$. On rejette H_0 et on déclare dans notre rapport que le nombre espèces de poissons est significativement plus faible en aval qu'en amont d'une décharge d'eaux usées. On peut également formuler la conclusion en déclarant que le nombre d'espèces de poissons est significativement plus élevée en amont qu'en aval d'une décharge

d'eaux usées dans notre aire d'étude. Finalement, on peut créer un graphique pour illustrer les résultats. Puisque les deux groupes sont appariés et que le test t pour deux groupes appariés est en fait un test t réalisé sur la différence entre les paires, on peut présenter la moyenne de la différence entre les paires et son intervalle de confiance à 95% (fig. 7).

```
> ##degrés de liberté résiduels
> res.df <- nrow(poissons) - 1
> ##IC à 95%
> inf95 <- mean.diff - qt(p = 0.025, df = res.df) * SE
> sup95 <- mean.diff + qt(p = 0.025, df = res.df) * SE
> ##graphique
> plot(x = 0.5, y = mean.diff,
       ylab = "Moyenne de différence entre les groupes",
       xlab = "Analyse réalisée sur le nombre d'espèces (aval - amont)",
       main = "Moyenne ± IC à 95%",
       ylim = range(c(inf95, sup95)),
       xaxt = "n")
> ##ajout d'IC à 95%
> segments(x0 = 0.5, x1 = 0.5,
           y0 = inf95, y1 = sup95)
```

Dans ce cas particulier, on remarque que la figure n'amène pas beaucoup plus d'information que de présenter la moyennne et l'intervalle de confiance directement dans le texte : $\bar{x}_{\text{Diff}} = -0.88$, IC à 95% : (-1.49, -0.26).

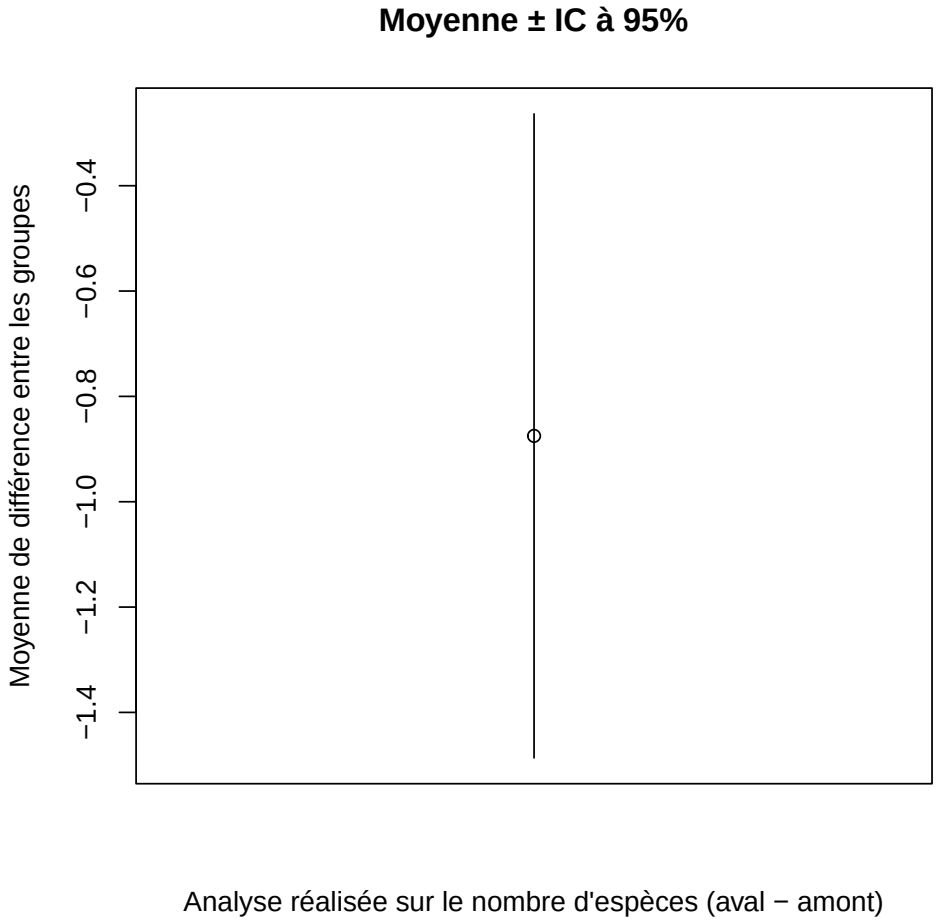

FIGURE 7 – Moyenne de la différence du nombre d'espèces de poissons (aval - amont) et intervalle de confiance à 95%.

4 Transformations

Parfois, les suppositions de normalité des résidus ou d'homogénéité de la variance ne peuvent pas être respectées. Dans de tels cas, il s'avère utile de transformer les données brutes. Dans cette section, nous verrons quelques transformations communes utilisées en statistiques.

4.1 Transformation logarithmique

La **transformation logarithmique** (*log transformation*) peut permettre d'arriver à la normalité ou d'homogénéiser les variances dans certains cas :

$$\begin{aligned} X' &= \log(X) \\ \text{ou} & \\ X' &= \log(X+1)\,, \end{aligned}$$

où X correspond à la variable originale et X' représente la nouvelle variable résultant de cette transformation sur chacune des valeurs originales. On ajoute une constante à cette transformation lorsque certaines des valeurs de la variable originale sont 0, puisque le logarithme n'est pas défini pour de telles valeurs. Bien qu'il soit possible d'utiliser n'importe quelle base pour le logarithme, la base e (log népérien, naturel) et la base 10 sont les plus fréquemment utilisées. En présence d'hétéroscédasticité, cette transformation peut rendre les variances homogènes (fig. 8). Cette transformation influence à la fois l'homogénéité de la variance et la normalité des résidus. Si les résidus ont déja une distribution normale, mais que les variances ne sont pas homogènes, il se peut que la transformation rétablisse l'homogénéité de la variance mais que la normalité des résidus ne soit plus rencontrée.

Dans `R`, on peut facilement appliquer la transformation logarithmique à l'aide de la fonction `log( )` sur une variable existante, comme celle du temps passé sur son téléphone vue plus tôt :

```
> ##transformation log sur Temps passé sur téléphone
> temps$log.Temps <- log(temps$Temps)
> head(temps)
```

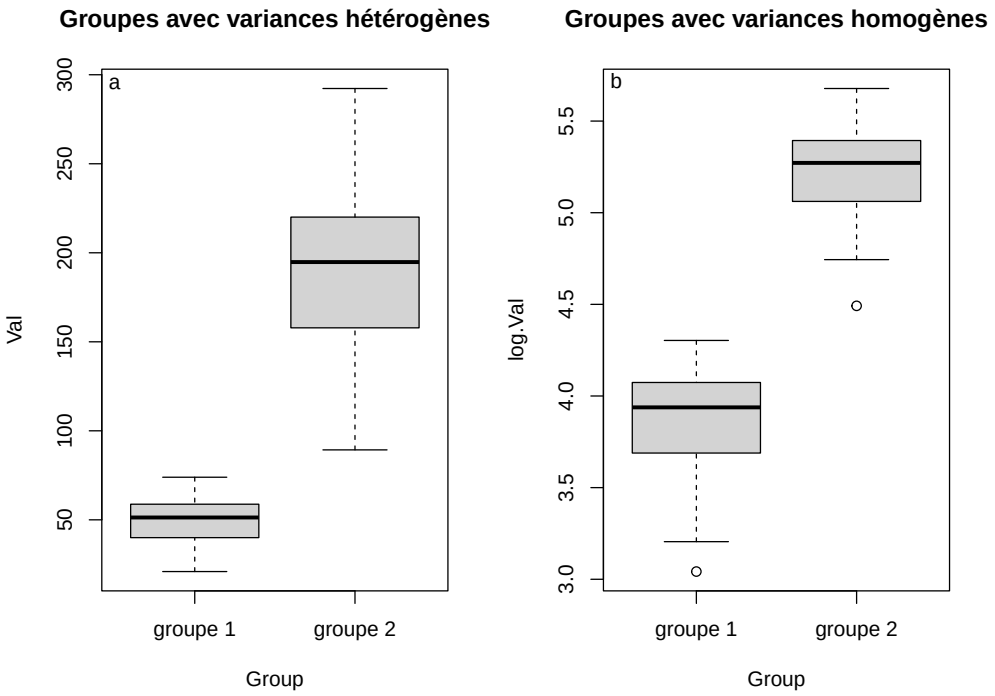

FIGURE 8 – Utilisation de la transformation logarithmique pour homogénéiser les variances de deux groupes. Les données brutes présentent des variances hétérogènes (a) puisque le groupe 1 varie moins que le groupe 2. L'application du log naturel a homogénéisé les variances (b), puisque la hauteur des deux boîtes est très semblable.

```
      Temps Accompagnateurs        Res log.Temps
1 27.05083        inconnus 11.2647400  3.297718
2 14.98468        inconnus -0.8014168  2.707028
3 26.18627        inconnus 10.4001760  3.265235
4 11.92027        inconnus -3.8658228  2.478240
5 16.49150        inconnus  0.7054111  2.802845
6 13.00014        inconnus -2.7859490  2.564960
```

4.2 Transformation racine carrée

La **transformation racine carrée** (*square-root transformation*) peut s'avérer utile lorsque les données sont sous forme de fréquences. Les fréquences sont constituées de valeurs entières (p. ex., 0, 10, 21) et ne peuvent prendre de valeurs décimales. Par exemple, dans une étude sur le nombre de naissances dans différents hôpitaux, on peut observer 3 naissances, ou encore 2 naissances, 1 naissances ou aucune, mais il est impossible d'observer 2.5 naissances dans un hôpital. De telles données suivent souvent une distribution autre que la distribution normale, comme la distribution de Poisson [2]). Avec des fréquences, les variances des groupes sont souvent proportionnelles aux moyennes [3] et la transformation peut homogénéiser les variances. Il existe différentes versions de la transformation racine carrée, chacune utilisant différentes constantes, notamment :

$$X' = \sqrt{X + 0.5}$$

et

$$X' = \sqrt{X + \frac{3}{8}}.$$

2. La distribution de Poisson est une distribution théorique qui définit les événements rares. Cette distribution ne possède qu'un seul paramètre estimé : la moyenne. Cette distribution théorique est utilisée dans certaines analyses, comme la régression de Poisson.

3. Rappelons que la variance et la moyenne sont deux paramètres indépendants dans la distribution normale, ce qui n'est pas le cas avec la distribution de Poisson.

La transformation racine carrée se réalise à l'aide de `sqrt( )` :

```
> ##on crée une variable de fréquences provenant
> ##d'une distribution de Poisson avec
> ##une moyenne de 2.5
> freqs <- rpois(n = 60, lambda = 2.5)
> ##transformation racine carrée sur des fréquences
> freqs.transf <- sqrt(freqs)
> ## On représente les deux distributions dans une même figure
> par(mfrow = c(1, 2))
> ## Distribution originale
> hist(freqs, ylim = c(0, 25))
> ## Distribution transformée
> hist(freqs.transf, ylim = c(0, 25))
```

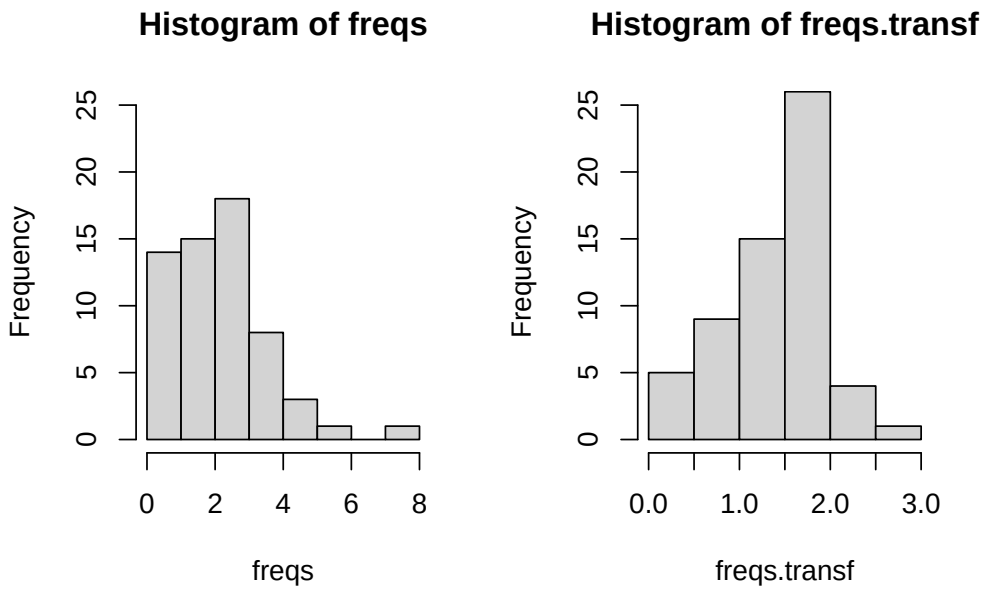

FIGURE 9 – Distribution de Poisson avec une moyenne de 2.5. Fréquences originales (gauche) et fréquences transformées suivant la transformation racine carrée (droite).

4.3 Transformation arcsinus

Les données sous formes de pourcentages et de proportions suivent fréquemment une distribution binomiale[4] plutôt que normale. Alors que la distribution normale est définie entre l'intervalle $[-\infty, +\infty]$, les pourcentages et les proportions sont limités dans l'intervalle $[0, 100]$ ou $[0, 1]$. Il en résulte que les données sous forme de pourcentages et de proportions ont plus de valeurs aux extrémités de l'intervalle (0 – 30 %, 70 – 100 %) qu'une distribution normale. Pour pallier ce problème, la **transformation arcsinus** (*arcsine transformation, angular transformation*) peut être utile :

$$X' = \arcsin \sqrt{X}$$

ou

$$X' = \sqrt{(n + \frac{1}{2})} \arcsin \sqrt{\frac{X + \frac{3}{8}}{n + \frac{3}{4}}} .$$

À noter qu'il faut ramener les pourcentages en proportions avant d'utiliser la transformation arcsinus (0.1 et 0.04, au lieu de 10 % et 4 %). Après la transformation, les données sont exprimées en radians ($\frac{180°}{2\pi}$). Dans `R`, on exécute la transformation arcsinus comme suit :

```
> ##on crée une série de proportions
> ##à partir d'une distribution bêta
> props <- rbeta(n = 60, shape1 = 5, shape2 = 1)
> ##transformation arcsinus sur les proportions
> props.transf <- asin(sqrt(props))
> ## On représente les deux distributions dans une même figure
> par(mfrow = c(1, 2))
> ## Distribution originale
> hist(props)
```

4. La distribution binomiale caractérise les événéments binaires, comme pile ou face, mort ou vivant, mâle ou femelle. Cette distribution est utilisée avec la régression logistique.

```
> ## Distribution transformée
> hist(props.transf)
```

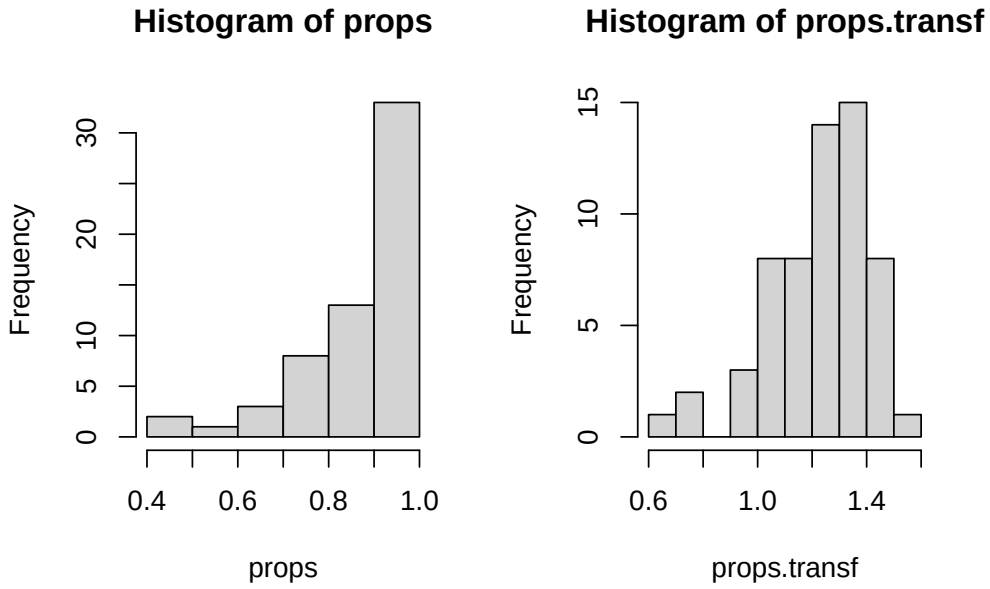

FIGURE 10 – Distribution bêta. Proportions originales (gauche) et proportions transformées suivant la transformation arcsinus (droite).

4.4 Autres transformations

La **transformation réciproque ou inverse** (*reciprocal transformation*) peut être utilisée lorsque les variances des groupes sont proportionnelles au carré des moyennes :

$$X' = \frac{1}{X}$$

ou

$$X' = \frac{1}{X+1}\,.$$

Toutefois, il faut rester vigilant dans l'interprétation des résultats lorsqu'on a appliqué la transformation inverse. Le signe sera inversé. Par exemple, si on obtient après la transforma-

tion inverse que la différence du groupe 1 est plus grande que celle du groupe 2, cela implique que pour les données brutes, c'est en effet le contraire qui se produit. La transformation inverse s'obtient de la manière suivante dans R :

```
> ##transformation inverse sur le Temps passé sur son téléphone
> temps$inv.Temps <- 1/(temps$Temps)
> head(temps)
     Temps Accompagnateurs        Res log.Temps  inv.Temps
1 27.05083        inconnus 11.2647400  3.297718 0.03696744
2 14.98468        inconnus -0.8014168  2.707028 0.06673484
3 26.18627        inconnus 10.4001760  3.265235 0.03818795
4 11.92027        inconnus -3.8658228  2.478240 0.08389072
5 16.49150        inconnus  0.7054111  2.802845 0.06063729
6 13.00014        inconnus -2.7859490  2.564960 0.07692223
```

La **transformation au carré** est parfois utile lorsque les variances diminuent alors que les moyennes augmentent, ou lorsque la distribution est allongée vers la gauche :

$$X' = X^2$$

On exécute la transformation au carré ainsi :

```
> ##transformation au carré sur le Temps passé sur son téléphone
> temps$carre.Temps <- (temps$Temps)^2
> head(temps)
     Temps Accompagnateurs        Res log.Temps  inv.Temps
1 27.05083        inconnus 11.2647400  3.297718 0.03696744
2 14.98468        inconnus -0.8014168  2.707028 0.06673484
3 26.18627        inconnus 10.4001760  3.265235 0.03818795
```

```
4 11.92027        inconnus -3.8658228  2.478240 0.08389072
5 16.49150        inconnus  0.7054111  2.802845 0.06063729
6 13.00014        inconnus -2.7859490  2.564960 0.07692223
  carre.Temps
1    731.7475
2    224.5405
3    685.7207
4    142.0928
5    271.9697
6    169.0037
```

Il existe d'autres familles de transformation afin de respecter les conditions d'utilisation des tests, mais celles que nous avons présentées plus haut sont les plus fréquemment utilisées. Malgré toutes ces possibilités de transformation des données, certaines variables ne peuvent être amenées à respecter les suppositions. Le cas type est celui où la variable contient une majorité de 0's.

Par exemple, lors d'une étude d'observation, on dénombre les arrestations pour vol à main armée dans 60 villages durant les 10 dernières années, et on observe que la majorité des villages n'ont eu aucun vol. Peu importe le type de transformation que nous utiliserons pour ce type de données, il y aura toujours plus de valeurs dans les queues de la distribution que ce qui est prédit par la distribution normale. Lorsqu'il est impossible de respecter les conditions d'utilisation d'un test d'hypothèse, il est préférable d'éviter de réaliser l'analyse et d'opter pour une analyse alternative. Ceci est particulièrement vrai lorsque la supposition d'homogénéité de la variance n'est pas respectée.

Conclusion

Dans ce texte, nous avons présenté des variantes du test t de Student pour comparer les données de deux groupes indépendants ainsi que pour des données appariées. Les suppositions sous-jacentes à ces analyses ont été énoncées, et plusieurs diagnostics formels et informels permettant de vérifier le respect de ces suppositions ont été illustrés. Nous avons également vu une modification du test t, celle de Welch, qui est appropriée lorsque les variances des groupes ne sont pas homogènes. Une série de transformations de données ont été présentées afin d'arriver à respecter les suppositions de normalité et d'homogénéité des variances.

Index